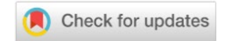
Advanced Engineering Research (Rostov-on-Don). 2024;24(4):413–423. eISSN 2687−1653

ИНФОРМАТИКА, ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА И УПРАВЛЕНИЕ
INFORMATION TECHNOLOGY, COMPUTER SCIENCE AND MANAGEMENT

УДК 519.688 *Оригинальное эмпирическое исследование*
https://doi.org/10.23947/2687-1653-2024-24-4-413-423

Алгоритм построения функции риска расширенной модели Кокса и его применение на базе данных больных раком предстательной железы


EDN: LNZDKF


И.И. Микулик[1] ✉, **Г.М. Жаринов**[2], **А.Ю. Кнеев**[2]

[1] Петербургский государственный университет путей сообщения Императора Александра I, г. Санкт-Петербург, Российская Федерация
[2] Российский научный центр радиологии и хирургических технологий имени академика А.М. Гранова Минздрава России, г. Санкт-Петербург, Российская Федерация
✉ mikulik.ilia@gmail.com

Аннотация
Введение. В медицине и связанных с нею отраслях для анализа выживаемости используются биоинспирированные подходы, среди которых особое место занимает регрессионная модель Кокса. Практика ее применения описана в теоретической и прикладной литературе. Однако требует тщательной проработки существенный недостаток данного метода. Дело в том, что признаки коррелируют с функцией риска линейно, и модель не задействует более сложные зависимости. Это создает трудности при исследовании анализа выживаемости. Представленная работа нацелена на решение данной проблемы. Объект изучения — расширенная модель Кокса, в которой функция риска включает нелинейную комбинацию признаков.
Материалы и методы. Использовалась база данных больных раком предстательной железы, так как в мировой онкологии это широко распространенный диагноз. Определен класс расширенных моделей Кокса с аддитивно-мультипликативной функцией риска. Для решения задачи методом оптимизации построена функция приспособленности, которая оценивает результаты прогнозов, количество признаков, а также степень переобучения модели — сложность и нагруженность составленной функции риска. Для оптимизации функции приспособленности разработан алгоритм муравьев-опылителей. Он имитирует размножение цветковых растений с помощью насекомых-опылителей и состоит из трех частей: муравьиный алгоритм, генетический алгоритм и алгоритм опыления. Качество обучения модели Кокса оценивали по С-индексу.
Результаты исследования. Предложен метаэвристический алгоритм оптимизации муравьев-опылителей, позволяющий строить функции риска расширенной модели Кокса. Набор параметров для обучения стандартной модели Кокса — весь используемый комплекс признаков: распространенность опухолевого процесса, время удвоения простатспецифического антигена (ПСА), сумма баллов по шкале Глисона, сывороточная концентрация ПСА на момент постановки диагноза, возраст и образование пациента, резус-фактор. Значение с-индекса обученной модели — 0,853691. Расширенная модель Кокса с найденной аддитивно-мультипликативной функцией риска имеет более высокий показатель С-индекса — 0,856241 с меньшим количеством используемых признаков (распространенность опухолевого процесса, время удвоения ПСА и сумма баллов по Глисону). По качеству этот подход не уступает классической модели Кокса или превосходит ее. Сокращение числа задействованных признаков должно повысить оперативность врачебного решения и ускорить начало лечения.
Обсуждение и заключение. Представленный алгоритм построения моделей анализа выживаемости повысил точность предсказания наступления терминального события и уменьшил количество используемых для этой цели признаков. Разница в точности для исследуемого набора данных представляется несущественной — С-индекс

возрос с 0,853691 до 0,856241 (на 0,3 %). При этом количество принимаемых во внимание признаков сократилось с 7 до 3 (на 57,1 %). Следовательно, предложенный метод эффективно решает задачу выбора признаков и может быть применен для повышения качества прогнозирования.

Ключевые слова: рак предстательной железы, прогнозирование выживаемости, вероятность наступления терминального события, регрессионная модель Кокса, алгоритм муравьев-опылителей

Благодарности. Авторы благодарят Благовещенскую Е.А., доктора физико-математических наук, профессора, за консультацию в области теории графов и алгоритмов оптимизации.

Для цитирования. Микулик И.И, Жаринов Г.М, Кнеев А.Ю. Алгоритм построения функции риска расширенной модели Кокса и его применение на базе данных больных раком предстательной железы. *Advanced Engineering Research (Rostov-on-Don)*. 2024;24(4):413–423. https://doi.org/10.23947/2687-1653-2024-24-4-413-423

Original Empirical Research

Algorithm for Constructing the Hazard Function of the Extended Cox Model and its Application to the Prostate Cancer Patient Database

Ilya I. Mikulik[1] ✉, **Gennadiy M. Zharinov**[2], **Aleksei Y. Kneev**[2]

[1] Emperor Alexander I St. Petersburg State Transport University, Saint Petersburg, Russian Federation

[2] Granov's Russian Research Center for Radiology and Surgical Technologies, Saint Petersburg, Russian Federation

✉ mikulik.ilia@gmail.com

Abstract

Introduction. In medicine and related industries, bioinspired approaches are used for the survival analysis, among which the Cox regression model holds a specific place. The practice of its application is described in the theoretical and applied literature. However, a significant drawback of this method requires careful study. The fact is that the features correlate with the hazard function linearly, and the model does not use more complex dependences. This causes some difficulties in studying survival analysis. The presented work is aimed at solving this problem. The object of study is the extended Cox model, in which the hazard function includes a nonlinear combination of features.

Materials and Methods. A database of prostate cancer patients was used, since this is a common diagnosis in global oncology. A class of extended Cox models with an additive/multiplicative hazard function was defined. To solve the problem using the optimization method, a fitness function was constructed that evaluated the results of prognosis, the number of features, and the degree of overtraining of the model — the complexity and load of the compiled hazard function. An algorithm of pollinating ants has been developed to optimize the fitness function. It simulates the reproduction of flowering plants using pollinating insects and consists of three parts: an ant colony algorithm, a genetic algorithm, and an ant pollinator algorithm. The quality of training of the Cox model was assessed by C-index.

Results. A metaheuristic algorithm for ant pollinator optimizing was proposed, providing for the construction of hazard functions of the extended Cox model. The set of parameters for training the standard Cox model was the entire set of features used: TNM, prostate-specific antigen doubling time (PSADT), Gleason score, serum PSA concentration at diagnosis, patient age and education, Rh factor. C-index value of the trained model was 0.853691. The extended Cox model with the found additive/multiplicative hazard function had a higher C-index value — 0.856241 with a smaller number of features used (TNM, PSADT, and Gleason score). In terms of quality, this approach is not inferior to or superior to the classical Cox model. Reducing the number of features involved should improve the efficiency of medical decisions and speed up the start of treatment.

Discussion and Conclusion. The presented algorithm for constructing survival analysis models increased the accuracy of predicting the occurrence of a terminal event, and reduced the number of features used for this purpose. The difference in accuracy for the studied data set seemed insignificant — C-index increased from 0.853691 to 0.856241 (by 0.3%). At this, the number of features taken into account was reduced from 7 to 3 (by 57.1%). Consequently, the proposed method effectively solves the problem of feature selection, and can be applied to improve the quality of prognostication.

Keywords: prostate cancer, survival prediction, terminal event probability, Cox regression model, ant pollinator algorithm

Acknowledgements. The authors would like to thank E.A. Blagoveshchenskaya, Dr.Sci. (Phys.-Math.), Professor, for consulting on issues of the graph theory and optimization algorithms.

For Citation. Mikulik II, Zharinov GM, Kneev AY. Algorithm for Constructing the Hazard Function of the Extended Cox Model and its Application to the Prostate Cancer Patient Database. *Advanced Engineering Research (Rostov-on-Don)*. 2024;24(4):413–423. https://doi.org/10.23947/2687-1653-2024-24-4-413-423

Введение. Анализ выживаемости представляет собой совокупность статистических методов, позволяющих оценить вероятность наступления терминального события, после которого объект выходит из-под наблюдения. Методы предполагают работу с данными, имеющими временную характеристику. Это время от начала наблюдения до наступления терминального события или выхода объекта из-под наблюдения. Возможность работы с объектами, вышедшими из-под наблюдения, представляет интерес для прикладных областей науки, в том числе для медицины [1].

Одна из классических моделей анализа выживаемости — регрессионная модель Кокса [2]. Ее функция риска использует линейную комбинацию признаков, что в общем случае может быть не вполне корректно, так как влияние признаков на значение функции риска может быть выражено нелинейной корреляцией. Для каждой задачи вклад признаков и функция риска могут коррелировать по-разному. Это определяется используемыми данными и требует особых подходов к поиску форм зависимостей. Разные способы определения зависимостей признаков в функции риска рассмотрены в [3]. В настоящей работе предлагается использовать расширенную модель Кокса, функция риска которой устанавливает не только аддитивную, но и мультипликативную комбинацию признаков. Кроме того, описан метод построения таких моделей в зависимости от используемых данных и набора признаков.

Построение модели предполагает решение задачи отбора признаков, одной из ключевых в анализе данных [4]. Она заключается в поиске оптимального набора признаков, достаточного для построения прогноза. Решение дает представление о том, какие признаки имеют бóльшую прогностическую значимость. Задачу можно сформулировать в терминах оптимизации и решить методами оптимизации. Предложенный для ее решения алгоритм муравьев-опылителей относится к метаэвристическим гибридным методам оптимизации. Он задействует муравьиный и генетический алгоритмы оптимизации, а также впервые разработанную модель скрещивания цветов.

Алгоритм реализован на базе данных больных раком предстательной железы. В мировой медицинской практике это одно из наиболее распространенных злокачественных новообразований у мужчин [5]. Внедрение скрининга на основе оценки сывороточной концентрации простатспецифического антигена (ПСА) существенно изменило структуру впервые выявленных случаев рака предстательной железы. Если ранее большинство из них приходилось на местно-распространенную и метастатическую формы опухоли, то в настоящее время доминирует локализованная. Благодаря этому увеличилась частота радикальных вмешательств и приблизились к 100 % показатели десятилетней выживаемости отдельных групп пациентов, перенесших радикальную простатэктомию или комбинированную гормонолучевую терапию.

Несмотря на очевидные успехи в диагностике и лечении рака предстательной железы, остаются нерешенными несколько важных вопросов, требующих исследования.

Современные методы прогнозирования выживаемости при раке предстательной железы основаны на совокупности факторов: возраст, распространенность и гистологическая дифференцировка опухоли, сывороточная концентрация ПСА, время его удвоения [6] и плотность [7]. Модель Кокса и другие модели анализа выживаемости дают о ней общее представление, но их точность в прогнозировании исходов для отдельных пациентов может варьироваться. Более того, прогноз, составленный по совокупности признаков, не дает представления о значимости каждого из них. Данное обстоятельство ограничивает возможности клиницистов адаптировать рекомендации по лечению к потребностям конкретного пациента.

Улучшение подходов к оценке выживаемости онкологического пациента — ключевой аспект научного поиска в области онкологии. Все больше внимания уделяется точности прогнозирования, которая критически важна для выбора терапевтической стратегии. Качественная прогностическая модель более точно определяет риск для больного и позволяет адаптировать подходы к лечению в зависимости от ожидаемого исхода. Это может улучшить и результаты лечения, и качество жизни пациента.

В условиях высокой нагрузки на медицинский персонал сокращение количества признаков в модели прогноза представляет значительную практическую ценность, так как сокращает временны́е затраты на принятие врачебных решений. Упрощение модели позволяет сделать акцент на ключевых аспектах клинической картины, что снижает вероятность некорректных интерпретаций данных. Кроме того, использование ограниченного набора признаков повышает воспроизводимость и стабильность результатов прогноза, то есть его надежность.

Цель настоящего исследования — разработка алгоритма построения моделей анализа выживаемости с отбором ключевых признаков. Точность нового подхода должна быть не ниже, чем у модели Кокса. Отметим, что различные способы построения функций риска модели Кокса задают не одну расширенную модель Кокса, а целый класс алгоритмов с различными функциями риска. Этот подход к адаптации функции риска под набор имеющихся данных и признаков выбран в качестве способа достижения поставленной цели.

Ниже перечислены задачи, решенные в данной работе.

1. Определен класс расширенных моделей Кокса с аддитивно-мультипликативной функцией риска.
2. Построена функция приспособленности, оценивающая результаты прогнозов расширенной модели Кокса.

3. Создан метод оптимизации, решающий поставленную задачу.

4. Разработана программа, реализующая предложенный алгоритм.

5. Получен результат работы программы на базе данных пациентов, больных раком предстательной железы, и показана эффективность разработанного алгоритма.

Материалы и методы. В анализе выживаемости для оценки риска наступления рассматриваемого события используются функции выживаемости и риска. Первая — это стохастическая характеристика, определяющая вероятность выживания (отсутствие терминального события) на протяжении заданного времени. Другими словами, функция выживаемости $S(t)$ определяется как вероятность того, что терминальное событие не наступит до момента времени t:

$$S(t) = P(T > t),$$

где T — время наступления терминального события.

Модели анализа выживаемости строят кривые выживаемости для каждого образца данных по его признакам. Модели часто задают с помощью функции риска, которая определяет вероятность наступления терминального события в бесконечно малый промежуток времени между t и Δt при условии, что оно не наступило до момента t:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P\left(t \le T < t + \Delta t \middle| T \ge t\right)}{\Delta t}.$$

Модель пропорциональных рисков Кокса вычисляет функцию риска для одного экземпляра как линейную комбинацию его признаков, устанавливая взаимосвязь между признаками экземпляра и функцией риска.

С одной стороны, явное задание функции риска делает модель прозрачной и удобной для интерпретации прогнозов. С другой стороны, предположение о линейной взаимосвязи признаков и прогноза является ограничением и не может выполняться для всех практических задач.

Результаты исследования. Пусть S — набор данных для обучения. Функция риска в классической модели Кокса:

$$\lambda\left(t \middle| X_i\right) = \lambda_0\, exp\left(\beta_1 X_{i1} + \ldots + \beta_p X_{ip}\right) = \lambda_0\, exp\left(\beta \cdot X_i\right),$$

где β — вектор влияния признаков; $X_i \in S$ — экземпляр данных.

В [3] функция риска модели Кокса рассматривается в обобщенном виде $\lambda(t|X_i) = \lambda_0 exp(g(\beta \cdot X_i))$, где $g(\beta \cdot X_i)$ — функция, устанавливающая зависимость между признаками экземпляров. В данной работе функция g является полиномом специального вида.

Пусть $F = \{f_1, f_2, \ldots, f_n\}$ — множество всех признаков. $|F| = p$. F_t — подмножество признаков: $F_t \subset F$. $P_q(F_t)$ — полином, составленный из признаков $f \in F_t$:

$$P_q(F_t) = \sum_{i=1}^{|\xi|} \varphi_i \sum_{j=1}^{|F_t|} f_i^{\xi_{ij}},$$

где $\xi = \{(\xi_1, \xi_2, \ldots, \xi_{|F_t|})\}$, $\xi_i \in \{0;\, 1\}$ — множество индикаторов вхождения i-го признака в слагаемое полинома; $\varphi_i \in \{0,\, 1\}$ — маркер, указывающий на вхождение i-го монома в P_q.

Таким образом, полином $P_q(F_t)$ — это сумма одночленов, каждый из которых является произведением признаков. При этом степень признака в одночлене не более единицы. Функцию риска $\lambda(t|X_i) = \lambda_0 exp(g(\beta \cdot X_i))$, где $g(\beta \cdot X_i) = P_q(F_t, \beta \cdot X_i)$, назовем аддитивно-мультипликативной, так как значение каждого признака в ней может входить либо в состав суммы, либо в состав произведения.

Для получения и обработки результатов необходимо оценить качество построенной модели. В работе [8] качество модели оценивается с помощью функции потерь, и это общий подход для любой обучающейся модели. В качестве оценивающего показателя чаще других используется индекс соответствия (С-индекс). Его же выбрали для оценки расширенной модели Кокса. С-индекс учитывает как наблюдаемые события, так и цензурированные случаи [9]. При этом количественно определяется ранговая корреляция между фактическим временем выживания и прогнозами модели. С-индекс показывает соотношение правильно упорядоченных (согласованных) и сопоставимых пар [10].

В работе рассматривается гипотеза о расширенной модели Кокса с точностью предсказаний не ниже, чем у классической модели Кокса.

Пусть $c(S, P_q(F_t))$ — С-индекс расширенной модели Кокса, обученной на данных S, с аддитивно-мультипликативной функцией риска, построенной с помощью полинома $P_q(F_t)$, а $c(S, P_\Sigma(F))$ — С-индекс классической модели Кокса, обученной на тех же данных S. Гипотезу можно сформулировать в виде:

$$\exists P_q(F_t), F_t \subset F : c\left(S, P_q(F_t)\right) \ge c\left(S, P_\Sigma(F)\right). \quad (1)$$

Для поиска нетривиальных примеров гипотезы поставили задачу и разработали алгоритм. Задачу можно сформулировать в терминах теории оптимизации. Необходимо построить полином $P_q(F_t)$ на подмножестве F с наибольшим значением $c(S, P_q(F_t))$ при минимальном наборе признаков F_t. Таким образом, вводятся два условия оптимизации $c(S, P_q(F_t)) \to max$ и $|F_t| \to min$. Следует учитывать еще одну проблему построения полинома. С увеличением числа возможных признаков экспоненциально растет количество возможных многочленов, в том числе сконструированных из-за переобучения. Как правило, такие многочлены состоят из суммы сравнительно большого количества мономов, а сами мономы — из большого количества множителей. Такие полиномы увеличивают точность модели лишь на обученных данных и слабо поддаются анализу.

Чтобы избавиться от проблемы переобучения, в работе предложена оптимизация по двум дополнительным критериям: количеству мономов в составе $P_q(F_t)$ и нагруженности полинома $P_q(F_t)$, которая отражает количество мультипликативных связей в полиноме.

Количество мономов определяется как $\sum_i^{|\xi|} \varphi_i$ Однако для построения корректного условия необходимо учитывать нелинейность вклада количества признаков в целевую функцию. При малом количестве входящих признаков ожидается более существенное изменение показателя, чем при большом. Поэтому в работе предложен показатель:

$$K_q = \frac{log_2\left(\sum_i^{|\xi|} \varphi_i\right)}{p}.$$

Отметим, что значение K_q не превышает 1.

$\sum_i^{|\xi|} \varphi_i \to max$ при $\forall\varphi_i = 1$, следовательно:

$$\sum_i^{|\xi|} \varphi_i = |\xi| = 2^p - 1,$$

$$log_2\left(2^p \text{ -} 1\right) < log_2 2^p = p,$$

$$K_q < 1.$$

Не совсем корректно определять нагруженность полинома как количество мультипликативных связей. В этом случае не отражается реальная оценка сложности полинома при разном количестве входящих в него мономов. Показатель должен демонстрировать нагруженность каждого входящего монома, поэтому в работе введена следующая величина, не превышающая 1:

$$B_q = \frac{\sum_{i,\varphi_i \neq 0} \sum_{j=1}^{|F|_t} \xi_{ij}}{\sum_i^{|\xi|} \varphi_i \cdot |F_t|}.$$

С учетом введенных характеристик задача оптимизации заключается в поиске $P_q(F_t)$ при условиях:

$$\begin{cases} c\left(S, P_q\left(F_t\right)\right) \to max, \\ F_t \to min, \\ K_q \to min, \\ B_q \to min. \end{cases} \quad (2)$$

Перейдем к одномерной оптимизации с помощью введения балансировочных коэффициентов ω:

$$f = \omega_1 c\left(S, P_q\left(F_t\right)\right) - \omega_2 \frac{F_t}{p} - \omega_3 K_q - \omega_4 B_q \to max.$$

Или представим в виде суммы:

$$f = \omega_1 \cdot c\left(S, P_q\left(F_t\right)\right) + \omega_2 \cdot \left(1 - \frac{F_t}{p}\right) + \omega_3 \cdot \left(1 - K_q\right) + \omega_4 \cdot \left(1 - B_q\right) \to max. \quad (3)$$

Последняя форма записи при необходимости позволяет зафиксировать значение целевой функции f, введя явную зависимость между балансировочными коэффициентами:

$$\omega_1 = \frac{\gamma_1}{1 + \gamma_2 + \gamma_3},$$

$$\omega_2 = \frac{\gamma_2}{1 + \gamma_1 + \gamma_3},$$

$$\omega_3 = \frac{\gamma_3}{1 + \gamma_1 + \gamma_2},$$

$$\omega_4 = 1 - \omega_1 - \omega_2 - \omega_3.$$

При любых $\gamma_1, \gamma_2, \gamma_3 \in (0; 1)$. Так, выбирая нужные γ_i или напрямую ω_i, можно усиливать или ослаблять соответствующие условия системы (2). Задача заключается в поиске максимума целевой функции f (3) при определенных ω_i.

Для решения задачи оптимизации в статье представлен разработанный алгоритм муравьев-опылителей. Он основан на модели муравьиной колонии, адаптированной под поставленную задачу. Алгоритм преобразовывает в модель набор вершин графа, представляющих признаки или их произведение. Имитируется процесс опыления и размножения цветковых растений с помощью насекомых-опылителей. Решение включает три алгоритма:

– муравьиный используется для построения модели;

– генетический улучшает работу муравьиного алгоритма;

– алгоритм опыления позволяет выбрать признаки или их произведения.

Результат работы алгоритма — полином $P_q(F_t)$, максимизирующий функцию f (3). Каждый моном, входящий в сумму полинома, представлен цветком. Множество цветков образует граф. По нему строят путь муравьи-опылители. Каждый муравей определяет множество цветов, и сумма соответствующих им мономов образует полином $P_q(F_t)$. Оценка построенного муравьем пути — это значение функции f (3) для расширенной модели Кокса с $g = P_q(F_t)$.

Муравьиный этап алгоритма представляет собой адаптированный к задаче простой муравьиный алгоритм [11]. Каждый муравей k имеет разный набор параметров α_k, β_k, Q_k. Чувствительность муравьев к феромонам α_k определяет степень эксплуатации муравьями найденных решений. Эвристическая чувствительность β_k устанавливает уровень эксплуатации эвристической информации. Интенсивность феромона Q_k определяет количество феромона, которое отложит муравей на цветок в процессе поиска решения. Статические параметры алгоритма: количество муравьев n, скорость испарения ρ, первоначальный уровень феромонов τ_0.

Каждый муравей выбирает вершину стохастически по правилу:

$$p_v^k(t) = \frac{\tau_v^{\alpha_k}(t)\eta_v^{\beta_k}}{\sum_u \tau_u^{\alpha_k}(t)\eta_v^{\beta_k}}, \quad (4)$$

где $p_v^k(t)$ — вероятность выбора цветка v муравьем k на итерации t; $\tau_v(t)$ — количество отложенного феромона на цветке v на итерации t; η_v — эвристическая информация, которая вычисляется как $\eta_v = c(S, P_i \equiv v)$. Во второй части этого равенства — c-индекс расширенной модели Кокса, обученной на одном мономе цветка v.

Каждый муравей откладывает феромон в соответствии с правилом:

$$\Delta\tau_v = \frac{Q_k}{f\left(P_q\left(F_t\right)\right)}, \quad (5)$$

где $P_q(F_t)$ — полином, построенный муравьем k; f — целевая функция.

Второй этап — приложение генетического алгоритма. Он модифицирует параметры муравьиного алгоритма с учетом эффективности найденных решений [12]. Алгоритм последовательно применяет к популяции муравьев (к их параметрам) три оператора: выбора, кроссинговера, мутации. В качестве оператора выбора используется метод рулетки. Муравей попадает в новую популяцию с вероятностью:

$$p_i = \frac{f\left(P_i\left(F_t\right)\right)}{\sum_j f\left(P_j\left(F_t\right)\right)}. \quad (5)$$

Оператор скрещивания — побитовая сумма битовых представлений параметров выбранных особей. Оператор мутации — инверсия случайного бита у битового представления параметра особи.

Феромоны, оставленные на вершинах-цветах, также применяются на этапе опыления. Данный этап использует популяционную идею. К популяции цветов прилагаются четыре оператора: селекции, кроссбридинга, лайнбридинга и старения. Каждый цветок кроме хранимого значения вершины-монома имеет параметр — возраст. Оператор селекции выбирает цветы с наибольшей концентрацией феромонов. Оператор кроссбридинга с некоторой вероятностью вводит новые цветы, моном которых представляет произведение объединения признаков из мономов цветов-родителей:

$$v_i = \left(e_i, \tau_i, \eta_i, o_i\right),$$

$$v_i \times v_j = v_k, v_k = \left(e_k = \prod f_q \in e_i \cup e_j\, f_q, \tau_k = \frac{\tau_i + \tau_j}{2}, \eta_k = c\left(S, P \equiv e_k\right), o_k = o_{max}\right),$$

где e — моном нового цветка; τ — случайное количество феромона, не превосходящее τ_0, отложенного на цветок; η — эвристическая составляющая; o — возраст цветка; o_{max} — установленная продолжительность жизни цветка.

Если в результате преобразования появились цветы, уже находящиеся в популяции, то новые цветы не создаются, а обновляется возраст у имеющихся. Оператор лайнбридинга с небольшой вероятностью добавляет в популяцию новый цветок с единственным признаком. Этот оператор используется, чтобы оставить возможность вытесненным признакам участвовать в работе алгоритма. Оператор старения понижает индикатор возраста у каждого цветка. Если индикатор старения стал равен нулю, цветок выбывает из популяции.

Таким образом, конфигурируемые параметры алгоритма: n, τ_0, ρ, o_{max}, α_0, β_0, Q_0, p_{kross}, p_{mut}. Выбор их значений зависит от текущей прикладной задачи и влияет на скорость сходимости алгоритма. Отметим, что параметры α_0, β_0, Q_0 адаптируются в ходе работы генетического алгоритма, поэтому их первоначальные значения не имеют большого влияния на скорость сходимости алгоритма, особенно при значительном количестве итераций представленного ниже алгоритма.

Начало

1. Определить параметры n, τ_0, ρ, o_{max}, α_0, β_0, Q_0, p_{kross}, p_{mut}

2. Положить $c = 0$, $P = \emptyset$

3. Положить множество цветов $V = \{v_i = (e_i = f_i,\ \tau_i = rand(0, \tau_0),\ \eta_i = c(S, P_i \equiv f_i),\ o_i = o_{max}) \mid \forall f_i \in F\}$

4. Положить множество муравьев $A = \{\alpha_k = (\alpha_k = \alpha_0,\ \beta_k = \beta_0,\ Q_k = Q_0)\}$

5. До достижения критерия остановки

5.1. Для каждого муравья $\alpha_k \in A$

5.1.1. $E_k(t) = \{v_{random}\}$

5.1.2. $c_k(t-1)$

5.1.3. $c_k(t) = \eta_i$

5.1.4. Пока $c_k(t) > c_k(t-1)$

5.1.4.1. Выбрать v в соответствии с правилом (4)

5.1.4.2. $E_k(t) = \cup\ \{v\}$

5.1.4.3. $c_k(t-1) = c_k(t)$

5.1.4.4. $P_k = \sum_{i, v_i \in E_k(t)} e_i$

5.1.4.5. $c_k(t) = f(S, P_k)$

5.1.5. Если $c_k(t) > c$

5.1.5.1. $c = c_k(t)$

5.1.5.2. $P = P_k$

5.1.6. Для каждого $v_i \in E_k(t)$ вычислить $\Delta\tau_v$ в соответствии с правилом (5)

5.2. Применить оператор выбора $A = S_{selection}(A)$

5.3. Применить оператор кроссинговера $A = S_{crossover}(A)$

5.4. Применить оператор мутации $A = S_{mutation}(A)$

5.5. Применить оператор селекции цветов $V = S_{selection}(V)$

5.6. Применить оператор кроссбридинга $V = S_{crossbreeding}(V)$

5.7. Применить оператор лайнбридинга $V = S_{linebreeding}(V)$

5.8. Применить оператор старения $V = S_{aging}(V)$

6. Вернуть значения c, P

Критерием остановки алгоритма может быть количество итераций или сходимость решений к одному значению. Таким образом, представленный метод оптимизации муравьев-опылителей решает задачу построения функции риска и отбора признаков для расширенной модели Кокса. Если для гипотезы (1) есть нетривиальные примеры, их можно найти описанным методом.

Алгоритм протестировали на базе данных больных раком предстательной железы. Они лечились или наблюдались с января 1996 по декабрь 2016 года в Российском научном центре радиологии и хирургических технологий имени академика А. М. Гранова Минздрава России [13]. В исследование включены обезличенные данные о распространенности опухолевого процесса у 5073 пациентов.

Перечень признаков, используемых в работе, с их описанием и количеством ненулевых записей представлен в таблице 1.

Таблица 1

Признаки набора данных

Название признака		Описание	Значение	Количество заполненных записей
краткое	полное			
‘ТР’	Тип распространения опухолевого процесса	Поражение соседних органов и структур, наличие регионарных и отдаленных метастазов	1 — локализованный 2 — местно-распространенный 3 — метастатический	5073
‘ВУ’	Время удвоения ПСА	Удвоение сывороточной концентрации ПСА, указывающее на возможное удвоение числа опухолевых клеток	Число с плавающей точкой	2423
‘ПГ’	Сумма баллов по шкале Глисона	Порядковая переменная. Отражает гистологическую дифференцировку опухоли	1 — ПГ < 7 2 — ПГ = 7 3 — ПГ > 7	3968
‘ПСА’	Сывороточная концентрация ПСА, послужившая основанием для биопсии	Простат-специфический антиген. Гликопротеин, сериновая протеаза в норме вырабатывается секреторным эпителием предстательной железы. Разжижает эякулят, улучшает подвижность сперматозоидов. Концентрация выше 4 нг/мл может быть основанием для биопсии	Число с плавающей точкой	4760
‘образование’	Уровень образования пациента	Завершенное образование пациента на момент постановки диагноза	0 — среднее общее 1 — среднее специальное 2 — высшее 3 — ученая степень	4622
‘возраст’	Возраст пациента	Возраст пациента на момент постановки диагноза	Целое число	5073
‘резус’	Резус-фактор	Наличие или отсутствие белка, отвечающего за резус-фактор	1 — положительный 2 — отрицательный	399

Не все признаки, представленные в таблице 1, существенны для исследования выживаемости. Есть и малозначимые (например, ‘образование’, ‘резус’). Они нужны для демонстрации корректной работы алгоритма, решающего задачу отбора признаков. Наличие коррелируемых и не очень важных признаков показывает практическую возможность использования алгоритма в условиях, когда заранее не известны ни зависимость признаков, ни их значимость.

Алгоритм реализован на языке программирования Python в пакете CoxPHFitter из библиотеки Lifelines. Для хранения и обработки данных использовалась программная библиотека Pandas.

Перед запуском алгоритма данные предварительно обработали. Это обусловлено тем, что в базе данных больных раком предстательной железы есть пропуски по ряду значений у некоторых пациентов. Для устранения проблемы использовали два способа обработки базы данных — удаление наблюдений и замена с учетом других значений в столбце [14]. Признаки ‘ТР’, ‘ВУ’ и ‘возраст’ являются важными и играют роль индикатора консистентности данных, поэтому удалялись наблюдения без этих признаков. Показатель Глисона ранжировали. Каждому наблюдению присвоили одно из трех значений: 1 — ПГ < 7 (1281 наблюдений); 2 — ПГ = 7 (1479 наблюдений); 3 — ПГ > 7 (1208 наблюдений).

Для остальных признаков отсутствующие значения заполняли методом *k*-взвешенных ближайших соседей. Такое восстановление пропущенных значений основывается на предположении, что близость экземпляров по

измеренным признакам указывает на их близость по неизмеренным признакам [15]. Метод *k*-взвешенных ближайших соседей предпочтителен ввиду небольших временных затрат на восстановление пропущенных значений [16], хотя есть более эффективные подходы [17].

Алгоритм реализован со следующим набором параметров: $n = 12$; $\tau_0 = 0{,}01$; $\rho = 0{,}8$; $o_{max} = 3$; $\alpha_0 = 0{,}5$; $\beta_0 = 2$; $Q_0 = 25$; $p_{kross} = 0{,}9$; $p_{mut} = 0{,}2$. Приведенный список значений параметров рекомендуется для первоначальной конфигурации алгоритма. Однако его можно изменить для решения конкретной задачи. В таблице 2 представлены результаты предложенного алгоритма.

Таблица 2

Значения С-индекса и функции приспособленности f в зависимости от полинома функции риска расширенной модели Кокса, найденного при заданных балансировочных коэффициентах

Полином аддитивно-мультипликативной функции риска расширенной модели Кокса	С-индекс	Функция приспособленности	Балансировочные коэффициенты ω_1; ω_2; ω_3; ω_4
‘ТР’+‘ВУ’	0,836789	0,782894	0,91;0,05;0,05;0,05
‘ТР’×‘ПГ’+‘ВУ’	0,840516	0,842814	0,99;0,05;0,05;0,05
‘ТР’+‘ВУ’+‘ПГ’	0,849790	0,746328	0,9;0,0;0,05;0,05
‘ТР’+‘ВУ’+‘ПГ’+‘ТР’×‘ВУ’	0,849828	0,827410	0,94;0,05;0,0;0,0
‘ТР’+‘ПГ’+‘ТР’×‘ВУ’×‘ПГ’	0,849830	0,841567	0,97;0,05;0,05;0,0
‘ТР’+‘ПГ’+‘ВУ’×‘ПГ’	0,850000	0,787661	0,94; 0,0;0,05;0,0
‘ТР’+‘ВУ’+‘ПГ’+‘ПСА’+‘образование’+ ‘возраст’+‘резус’	0,853691	0,838012	0,99; 0,0;0,0;0,05
‘ТР’+‘ВУ’+‘ПГ’+‘ТР’×‘ПГ’	0,855292	0,809308	0,94;0,05;0,0;0,05
‘ТР’+‘ТР’×‘ПГ’+‘ПГ’+‘ТР’×‘ВУ’	0,856241	0,764870	0,91;0,0;0,05;0,0
‘ТР’+‘ТР’×‘ПГ’×‘ПСА’+‘ПГ’+‘ПСА’+ ‘ПГ’×‘ПСА’+‘ВУ’+‘ТР’×‘ПГ’+‘ТР’×‘ПСА’	0,861085	0,839459	0,95;0,05;0,0;0,0
‘ТР’+‘ВУ’+‘ПГ’+‘ПСА’+‘образование’+ ‘возраст’+‘резус’+‘ТР’×‘ПГ’+‘ТР’×‘ПСА’+ ‘ПГ’×‘ПСА’	0,861643	0,826508	0,97;0,0;0,0;0,05
‘ТР’+‘ВУ’+‘ПГ’+‘ПСА’+‘образование’+ ‘резус’+‘ТР’×‘ПГ’+‘ТР’×‘ПСА’+‘ПГ’× ‘ПСА’+‘ПСА’×‘возраст’	0,862345	0,845098	0,98;0,0;0,0;0,0

В таблице 2 вариации вхождения признаков в функцию риска ранжированы по возрастанию С-индекса. Здесь также указаны значения балансировочных коэффициентов фитнесс-функции (3), при которых найдено представленное решение и значение самой функции. Последние строки таблицы содержат функции риска с наиболее высоким индексом согласованности. Они достаточно сложны для анализа из-за нагруженности, связанной с низкими значениями соответствующих балансировочных коэффициентов.

Обсуждение и заключение. Лучший набор признаков для обучения стандартной (нерасширенной) модели Кокса — это весь представленный набор признаков, то есть функция ‘ТР’ + ‘ВУ’ + ‘ПГ’ + ‘ПСА’ + ‘образование’ + ‘возраст’ + ‘резус’ со значением с-индекса 0,853691. В то же время расширенная модель Кокса с найденной функцией риска ‘ТР’ + ‘ТР’ × ‘ПГ’ + ‘ПГ’ + ‘ТР’ × ‘ВУ’ имеет более высокий показатель с-индекса — 0,856241 с меньшим количеством используемых признаков.

Итоги данной научной работы позволяют сделать определенные выводы. Если иметь в виду представленную базу данных, то параметров ‘ТР’, ‘ВУ’, ‘ПГ’ достаточно для построения качественной модели анализа выживаемости. Таким образом, результат исследования — это возможность построения модели выживаемости с меньшим количеством используемых признаков. Причем предложенное решение не уступает или превосходит результативность классической модели Кокса, для обучения которой задействуют много признаков.

Алгоритм, созданный в рамках данной работы, способен решать задачу нахождения лучшей комбинации признаков за приемлемое число итераций (30). Набор регуляризующих коэффициентов позволяет задать алгоритму определенную конфигурацию. Благодаря этому специалист прикладной области может сделать выбор в пользу улучшения качества предсказания, сокращения количества признаков или исключения проблемы переобучения.

Итак, класс метаэвристических алгоритмов приемлем для решения поставленной задачи. На этапе опыления строятся мономы в полиноме, то есть проводится поиск мультипликативных зависимостей признаков. На этапе муравьиного алгоритма строится полином из мономов, то есть идет поиск аддитивных зависимостей признаков. Генетический этап необходим, чтобы улучшить сходимость и стабильность работы муравьиного алгоритма.

Для рассмотренного набора данных предложенный алгоритм повысил точность предсказания. Правда, незначительно. С-индекс увеличился всего на 0,3 %, с 0,853691 до 0,856241. Однако количество рассматриваемых признаков сократилось на 57,1 %, с 7 до 3. Меньшее число признаков в прогностической модели облегчает работу врачей, позволяет выиграть время при принятии решений и может снизить вероятность ошибок при интерпретации данных.

Список литературы / References

1. Archetti A, Lomurno E, Lattari F, Martin A, Matteucci M. Heterogeneous Datasets for Federated Survival Analysis Simulation. In: *Proc. Companion of the 2023 ACM/SPEC International Conference on Performance Engineering*. New York: Association for Computing Machinery; 2023. P. 173–180. http://doi.org/10.1145/3578245.3584935

2. Atlam M, Torkey H, El-Fishawy N, Salem H. Coronavirus Disease 2019 (COVID-19): Survival Analysis Using Deep Learning and Cox Regression Model. *Pattern Analysis and Applications*. 2021;24:993–1005. http://doi.org/10.1007/s10044-021-00958-0

3. Govindarajulu US, Malloy EJ, Ganguli B, Spiegelman D, Eisen EA. The Comparison of Alternative Smoothing Methods for Fitting Non-Linear Exposure-Response Relationships with Cox Models in a Simulation Study. *The International Journal of Biostatistics*. 2009;5(1):2. http://doi.org/ 10.2202/1557-4679.1104

4. Miren Hayet-Otero, Fernando García-García, Dae-Jin Lee, Joaquín Martínez-Minaya, Pedro Pablo España Yandiola, Isabel Urrutia Landa, et al. Extracting Relevant Predictive Variables for COVID-19 Severity Prognosis: An Exhaustive Comparison of Feature Selection Techniques. *PLoS One.* 2023;18(4):e0284150. https://doi.org/10.1371/journal.pone.0284150

5. Berenguer CV, Pereira F, Câmara JS, Pereira JA. Underlying Features of Prostate Cancer — Statistics, Risk Factors, and Emerging Methods for Its Diagnosis. *Current Oncology.* 2023;30(2):2300–2321. https://doi.org/10.3390/curroncol30020178

6. Жаринов, Г.М., Богомолов О.А. Исходное время удвоения простатспецифического антигена: клиническое и прогностическое значение у больных раком предстательной железы. *Онкоурология.* 2014;(1):44–48.

Zharinov GM, Bogomolov OA. The Pretreatment Prostate-Specific Antigen Doubling Time: Clinical and Prognostic Values in Patients with Prostate Cancer. *Cancer Urology.* 2014;(1):44–48.

7. Kneev AY, Shkol'nik MI, Bogomolov OA, Zharinov GM. Prostate Specific Antigen Density as a Prognostic Factor in Patients with Prostate Cancer Treated with Combined Hormonal Radiation Therapy. *Siberian Journal of Oncology.* 2022;21(3):12–23. https://doi.org/10.21294/1814-4861-2022-21-3-12-23

8. Ewees AA, Al-qaness MA Abualigah L, Oliva D, Algamal ZY, Anter AM, et al. Boosting Arithmetic Optimization Algorithm with Genetic Algorithm Operators for Feature Selection: Case Study on Cox Proportional Hazards Model. *Mathematics*. 2021;9(18):2321. https://doi.org/10.3390/math9182321

9. Alabdallah A, Ohlsson M, Pashami S, Rögnvaldsson Th. The Concordance Index Decomposition: A Measure for a Deeper Understanding of Survival Prediction Models. *Artificial Intelligence in Medicine*. 2024;148:102781. https://doi.org/10.48550/ARXIV.2203.00144

10. Cavalcante Th, Ospina R, Leiva V, Cabezas X, Martin-Barreiro C. Weibull Regression and Machine Learning Survival Models: Methodology, Comparison, and Application to Biomedical Data Related to Cardiac Surgery. *Biology.* 2023;12(3):442. https://doi.org/10.3390/biology12030442

11. Guangyu Liu, Yuwei Bai, Ling Zhu, Qingyun Wang, Wei Zhang. A Sequential Excitation and Simplified Ant Colony Optimization Based Global Extreme Seeking Control Method for Performance Improvement. *Swarm and Evolutionary Computation*. 2024;86:101522. https://doi.org/10.1016/j.swevo.2024.101522

12. Blagoveshchenskaya EA, Mikulik II, Strüngmann LH. Ant Colony Optimization with Parameter Update Using a Genetic Algorithm for Travelling Salesman Problem. In: *Proc. Workshop "Models and Methods for Researching Information Systems in Transport"*. 2020;2803:20–25. URL: https://ceur-ws.org/Vol-2803/paper3.pdf (accessed: 17.09.24).

13. Жаринов Г.М. *База данных больных раком предстательной железы*. База данных РФ. № 2016620331. 2016. 1 с. URL: https://www1.fips.ru/fips_servl/fips_servlet?DB=DB&DocNumber=2016620331&TypeFile=html (дата обращения: 17.09.2024).

Zharinov GM. Prostate Cancer Patients Database. RF Database, no. 2016620331. 2016. 1 p. (in Russ.) URL: https://www1.fips.ru/fips_servl/fips_servlet?DB=DB&DocNumber=2016620331&TypeFile=html (accessed: 17.09.2024).

14. Ghannad-Rezaie M, Soltanian-Zadeh H, Hao Ying, Ming Dong. Selection-Fusion Approach for Classification of Datasets with Missing Values. *Pattern Recognition*. 2010;43(6):2340–2350. https://doi.org/10.1016/j.patcog.2009.12.003

15. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing Value Estimation Methods for DNA Microarrays. *Bioinformatics*. 2001;17(6):520–525. https://doi.org/10.1093/bioinformatics/17.6.520

16. Koshechkin AA, Andryushchenko VS, Zamyatin AV. A New Method to Missing Value Imputation for Immunosignature Data. *CTM* (*Sovremennye tehnologii v medicine)*. 2019;11(2):19–24. https://doi.org/10.17691/stm2019.11.2.03

17. Eunseo Oh, Hyunsoo Lee. Quantum Mechanics-Based Missing Value Estimation Framework for Industrial Data. *Expert Systems with Applications.* 2024;236:121385. https://doi.org/10.1016/j.eswa.2023.121385

Об авторах:

Илья Игоревич Микулик, аспирант кафедры высшей математики Петербургского государственного университета путей сообщения Императора Александра I (190031, Российская Федерация, г. Санкт-Петербург, Московский пр., 9), SPIN-код, ORCID, ScopusID, ResearcherID, mikulik.ilia@gmail.com

Геннадий Михайлович Жаринов, доктор медицинских наук, профессор, главный научный сотрудник отдела лучевых и комбинированных методов лечения РНЦРХТ им. акад. А.М. Гранова Минздрава России (197758, Российская Федерация, г. Санкт-Петербург, пос. Песочный, ул. Ленинградская, 70), SPIN-код, ORCID

Алексей Юрьевич Кнеев, кандидат медицинских наук, старший преподаватель кафедры радиологии, хирургии и онкологии, врач-онколог отделения онкоурологии РНЦРХТ им. акад. А.М. Гранова Минздрава России (197758, Российская Федерация, г. Санкт-Петербург, пос. Песочный, ул. Ленинградская, 70), SPIN-код, ORCID, ScopusID

Заявленный вклад авторов:

И.И. Микулик: разработка и реализация метода исследования — алгоритма муравьев-опылителей.

Г.М. Жаринов: постановка цели исследования, предоставление базы данных для исследования, описание характеристик базы данных, описание прикладной задачи.

А.Ю. Кнеев: описание актуальности и результатов исследования.

Конфликт интересов: **авторы заявляют об отсутствии конфликта интересов.**

Все авторы прочитали и одобрили окончательный вариант рукописи.

About the Authors:

Ilya I. Mikulik, Postgraduate student of the Higher Mathematics Department, Emperor Alexander I St. Petersburg State Transport University (9, Moskovsky Pr., St. Petersburg, 190031, Russian Federation), SPIN-code, ORCID, ScopusID, ResearcherID, mikulik.ilia@gmail.com

Gennadiy M. Zharinov, Dr.Sci.(Medicine), Professor, Chief Researcher of the Department of Radiation and Combined Methods of Treatment, Granov's Russian Research Center for Radiology and Surgical Technologies (70, Leningradskaya Str., v. Pesochny, St. Petersburg, 197758, Russian Federation), SPIN-code, ORCID

Aleksei Yu. Kneev, Cand.Sci.(Medicine), Senior Lecturer of the Department of Radiology, Surgery and Oncology, Oncologist of the Department of Oncourology, Granov's Russian Research Center for Radiology and Surgical Technologies (70, Leningradskaya Str., v. Pesochny, St. Petersburg, 197758, Russian Federation), SPIN-code, ORCID, ScopusID

Claimed Contributorship:

II Mikulik: development and implementation of the research method — the ant pollinator algorithm.

GM Zharinov: setting the research objective, providing a database for research, describing the characteristics of the database, describing the applied task.

AY Kneev: description of the urgency of the study and the research results.

Conflict of Interest Statement: **the authors declare no conflict of interest.**

All authors have read and approved the final manuscript.

Поступила в редакцию / Received 28.10.2024
Поступила после рецензирования / Reviewed 22.11.2024
Принята к публикации / Accepted 02.12.2024